

Distributed Statistical Estimation and Rates of Convergence in Normal Approximation

Stanislav Minsker

E-mail: minsker@usc.edu

Nate Strawn

E-mail: nate.strawn@georgetown.edu

Summary. This paper presents new algorithms for distributed statistical estimation that can take advantage of the divide-and-conquer approach. We show that one of the key benefits attained by an appropriate divide-and-conquer strategy is robustness, an important characteristic of large distributed systems. We introduce a class of algorithms that are based on the properties of the geometric median, establish connections between performance of these distributed algorithms and rates of convergence in normal approximation, and provide tight deviations guarantees for resulting estimators in the form of exponential concentration inequalities.

Our techniques are illustrated through several examples: in particular, we obtain new results for the median-of-means estimator, as well as provide performance guarantees for robust distributed maximum likelihood estimation.

1. Introduction.

According to (IBM, 2015), “Every day, we create 2.5 quintillion bytes of data so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos.. to name a few. This data is big data”. Novel scalable and robust algorithms are required to successfully address the challenges posed by big data problems. This paper develops and analyzes techniques that exhibit *scalability*, a necessary characteristic of modern methods designed to perform statistical analysis of large datasets, as well as *robustness* that guarantees stable performance of distributed systems when some of the nodes exhibit abnormal behavior.

The computational power of a single computer is often insufficient to store and process modern data sets, and instead data is stored and analyzed in a distributed way by a cluster consisting of several machines. We consider a distributed estimation framework wherein data is assumed to be randomly assigned to computational nodes that produce intermediate results. We assume that no communication between the nodes is allowed at this first stage. On the second stage, these intermediate results are used to compute some statistic on the whole dataset; see figure 1 for a graphical illustration. Often, such a distributed setting is unavoidable in applications, whence interactions between sub-samples stored on different machines are inevitably lost. Most previous research focused on the following question: how significantly does this loss affect the quality of statistical estimation when compared to an “oracle” that has access to the whole sample?

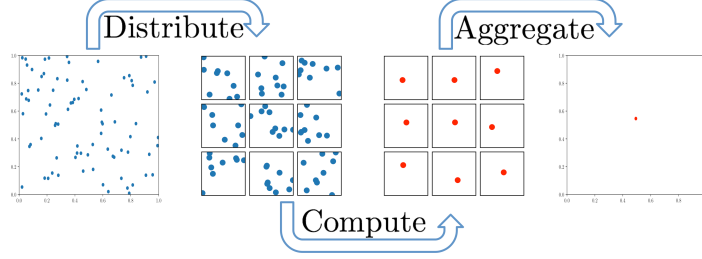


Fig. 1: Distributed estimation protocol where data is randomly distributed across nodes to obtain “local” estimates that are aggregated to compute a “global” estimate.

The question that we ask in this paper is different: what can be gained from randomly splitting the data across several subsamples? What are the statistical advantages of the divide-and-conquer framework? Our work indicates that one of the key benefits of an appropriate merging strategy is robustness. In particular, the quality of estimation attained by the distributed estimation algorithm is preserved even if a subset of machines stops working properly. At the same time, the resulting estimators admit tight probabilistic guarantees (expressed in the form of exponential concentration inequalities) even when the distribution of the data has heavy tails – a viable model of real-world samples contaminated by outliers.

We establish connections between a class of randomized divide-and-conquer strategies and the rates of convergence in normal approximation. Using these connections, we provide a new analysis of the “median-of-means” estimator which often yields significant improvements over the previously available results. We further illustrate the implications of our results by constructing novel algorithms for distributed Maximum Likelihood Estimation that admit strong performance guarantees under weak assumptions on the underlying distribution.

1.1. Background and related work.

We begin by introducing a simple model for distributed statistical estimation. Let X be a random variable taking values in a measurable space (S, \mathcal{S}) , and let P be the distribution of X . Moreover, let X_1, \dots, X_N be a sequence of i.i.d. copies of X representing the data available to a statistician. We will assume that N is large, and that the sample $\mathcal{X} = (X_1, \dots, X_N)$ is *randomly partitioned* into k disjoint subsets G_1, \dots, G_k of cardinality $n \geq \lfloor \frac{N}{k} \rfloor$ each. The goal is to estimate an unknown parameter $\theta_* = \theta_*(P)$ taking values in a separable Hilbert space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ (for example, if $S = \mathbb{H}$, it could be the mean of X) by performing “local” computations with each subset G_j , $j \leq k$. The local estimators $\bar{\theta}_j := \bar{\theta}_j(G_j)$, $j \leq k$ are then pieced together to produce the final estimator $\hat{\theta}^{(k)} = \hat{\theta}^{(k)}(\bar{\theta}_1, \dots, \bar{\theta}_k)$. We are interested in the statistical properties of this distributed estimation protocol, and our main focus is on the final step that combines the local estimators. Let us mention that the condition requiring the sets G_j , $1 \leq j \leq k$ to be disjoint can be relaxed; we discuss the alternative possibilities in section 3.2 below.

The problem of distributed and communication - efficient statistical estimation has recently received significant attention from the research community. While our review

provides only a subsample of the abundant literature in this field, it is important to acknowledge the works by McDonald et al. (2009); Zhang et al. (2012); Fan et al. (2014); Battey et al. (2015); Duchi et al. (2014); Lee et al. (2015); Cheng and Shang (2015); Rosenblatt and Nadler (2016); Zinkevich et al. (2010). Li et al. (2016); Scott et al. (2016); Shang and Cheng (2015); Minsker et al. (2014) have investigated closely related problems for distributed Bayesian inference. Applications to important algorithms such as Principal Component Analysis were investigated in (Fan et al., 2017; Liang et al., 2014), among others. Jordan (2013), author provides an overview of recent trends in the intersection of the statistics and computer science communities, describes popular existing strategies such as the “bag of little bootstraps”, as wells as successful applications of the divide-and-conquer paradigm to problems such as matrix factorization.

The majority of the aforementioned works propose *averaging* of local estimators as a final merging step. Indeed, averaging reduces variance, hence, if the bias of each local estimator is sufficiently small, their average often attains optimal rates of convergence to the unknown parameter θ_* . For example, when $\theta_*(P) = \mathbb{E}_P X$ is the mean of X and $\bar{\theta}_j$ is the sample mean evaluated over the subsample G_j , $j = 1, \dots, k$, then the average of local estimators $\tilde{\theta} = \frac{1}{k} \sum_{j=1}^k \bar{\theta}_j$ is just a empirical mean evaluated over the whole sample. More generally, it has been shown by Battey et al. (2015); Zhang et al. (2013) that in many problems (for instance, linear regression), k can be taken as large as $O(\sqrt{N})$ without negatively affecting the estimation rates; similar guarantees hold for a variety of M-estimators (see Rosenblatt and Nadler, 2016). However, if the number of nodes k itself is large (the case we are mainly interested in), then the averaging scheme has a significant drawback. In a real-world scenarios, it is common for one or more of the local estimators $\bar{\theta}_j$ to be anomalous, for example, due to data corruption or computer system malfunctioning, whence statistical properties of the average will be negatively affected as well. For large distributed systems, this drawback can be costly.

One way to address this issue is to replace averaging by a more robust procedure, such as the median; this is the approach we take in the present work. Specifically, we set

$$\hat{\theta}^{(k)} = \text{med}(\bar{\theta}_1, \dots, \bar{\theta}_k)$$

for an appropriately defined median, for example the *geometric median* (Small, 1990). Since the median remains stable as long as at least a half of the nodes in the system perform as expected, such model for distributed estimation is robust. This approach has been proposed and investigated by Minsker (2015) and Hsu and Sabato (2016).

However, *existing results* for the median-based approach have several pitfalls related to the convergence rates, and in most cases known guarantees are suboptimal. In particular, these guarantees suggest that estimators obtained via the median-based approach are very sensitive to the choice of k , the number of partitions. For instance, consider the problem of univariate mean estimation: $X \in \mathbb{R}$, $\theta_* = \mathbb{E}X$ is the expectation of X . Let $\bar{\theta}_j = \frac{1}{|G_j|} \sum_{i: X_i \in G_j} X_i$ be the empirical mean evaluated over the subsample G_j , $j = 1, \dots, k$, and define the “median-of-means” estimator via

$$\hat{\theta}^{(k)} = \text{med}(\bar{\theta}_1, \dots, \bar{\theta}_k), \quad (1)$$

where $\text{med}(\cdot)$ is the usual univariate median. This estimator has been introduced by Nemirovski and Yudin (1983) in the context of stochastic optimization, and later ap-

peared in (Jerrum et al., 1986) and (Alon et al., 1996). If $\text{Var}(X) = \sigma^2 < \infty$, it has been shown (for example, by Lerasle and Oliveira, 2011) that the median-of-means estimator $\widehat{\theta}^{(k)}$ satisfies

$$|\widehat{\theta}^{(k)} - \theta_*| \leq 2\sigma\sqrt{6e}\sqrt{\frac{k}{N}} \quad (2)$$

with probability $\geq 1 - e^{-k}$. However, this bound, while being the current state of the art, does not tell us what happens at the confidence levels other than $1 - e^{-k}$. For example, if $k = \lfloor \sqrt{N} \rfloor$, the only conclusion we can make is that $|\widehat{\theta}^{(k)} - \theta_*| \lesssim N^{-1/4}$ with high probability, which is far from the optimal rate $N^{-1/2}$. And if we want the bound to hold with confidence 99% instead of $1 - e^{-\sqrt{N}}$, then, according to (2), we should take $k = \lfloor \log 100 \rfloor + 1 = 5$, in which case the beneficial effect of parallel computation is very limited. The natural question to ask is the following: is the median-based merging step indeed suboptimal for large values of k (e.g., $k = \lfloor \sqrt{N} \rfloor$), or is the problem related to the suboptimality of existing bounds? We claim that in many situations the latter is the case, and that previously known results can be strengthened: for instance, the statement of Corollary 1 below implies that whenever $\mathbb{E}|X - \theta_*|^3 < \infty$, the median-of-means estimator satisfies

$$|\widehat{\theta}^{(k)} - \theta_*| \leq 3\sigma \left(\frac{\mathbb{E}|X - \theta_*|^3}{\sigma^3} \frac{k}{N - k} + \sqrt{\frac{s}{N - k}} \right)$$

with probability $\geq 1 - 4e^{-2s}$, for all $s \lesssim k$. In particular, this inequality shows that estimator (1) attains optimal rate $N^{-1/2}$ whenever $k = O(\sqrt{N})$, hence the “statistical cost” of employing a large number of computational nodes is minor.

We demonstrate that improved bounds hold in other important scenarios (such as maximum likelihood estimation) as well, and require only weak assumptions on the underlying distribution.

1.2. Organization of the paper.

Section 2 describes notation used throughout the paper. Section 3 introduces our main results which are illustrated by several examples. Bounds in the univariate case (that is, $\theta_* \in \mathbb{R}$) admit sharper constants, so we present them separately. The proofs of main results are contained in section 4.

2. Notation.

Everywhere below, $\|\cdot\|_2$ stands for the Euclidean norm of a vector and $\|\cdot\|$ - for the operator norm of a matrix (it largest singular value). Condition number $\text{cond}(A)$ of a non-singular matrix A is defined as $\text{cond}(A) = \|A\| \|A^{-1}\|$.

Given a probability measure P , $\mathbb{E}_P(\cdot)$ will stand for the expectation with respect to P , and we will write $\mathbb{E}(\cdot)$ when P is clear from the context. At times, it will be more convenient to use $Pf := \mathbb{E}_P f(X)$ to denote the expectation of $f(X)$, where P is the distribution of X .

For two sequences $\{a_j\}_{j \geq 1} \subset \mathbb{R}$ and $\{b_j\}_{j \geq 1} \subset \mathbb{R}$ for $j \in \mathbb{N}$, the expression $a_j \lesssim b_j$ means that there exists a constant $c > 0$ such that $a_j \leq cb_j$ for all $j \in \mathbb{N}$. Finally, for a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we define

$$\operatorname{argmin}_{z \in \mathbb{R}^d} f(z) = \{z \in \mathbb{R}^d : f(z) \leq f(x) \text{ for all } x \in \mathbb{R}^d\}.$$

Additional notation and auxiliary results are introduced on demand for the proofs in section 4.

3. Main results.

As we have argued above, existing guarantees for the estimator (1) are very sensitive to the choice of k , the number of partitions. In this section, we demonstrate that these bounds are often suboptimal, and show that large values of k often have negligible impact on the performance of resulting algorithm.

The key observation underlying the subsequent exposition is the following: assume that the “local estimators” $\bar{\theta}_j$, $1 \leq j \leq k$, are identically distributed and unbiased (or have “small” bias). Moreover, suppose that their common distribution \bar{P}_k is approximately symmetric (that is, the laws of $\bar{\theta}_j$ and $-\bar{\theta}_j$ are “close”). It implies that the distance between the median of \bar{P}_k and its mean θ_* is small, hence the sample median evaluated over the i.i.d. sample $W_1 = \bar{\theta}_1, \dots, W_k = \bar{\theta}_k$ must also be close (with high probability) to the unknown θ_* . For instance, in the mean estimation example, “approximate symmetry” follows from the Central Limit Theorem. We formalize this intuition below. Results for the univariate case (with sharper constants) are presented in section 3.1, and extensions to the multivariate case are given in section 3.3.

3.1. The univariate case.

Let $X \in \mathbb{R}^d$ be a random vector with distribution P , and let $\theta_* = \theta_*(P) \in \mathbb{R}$ be a real-valued parameter of interest. As before, we assume that X_1, \dots, X_N is a collection of i.i.d. copies of X that is randomly partitioned into disjoint groups G_1, \dots, G_k of cardinality $n = \lfloor N/k \rfloor$ each (equality of group sizes is not a requirement but a technically convenient assumption). Let $\bar{\theta}_j := \bar{\theta}_j(G_j)$, $1 \leq j \leq k$ be a sequence of i.i.d. estimators of θ_* and $\hat{\theta}^{(k)} = \operatorname{med}(\bar{\theta}_1, \dots, \bar{\theta}_k)$. Moreover, suppose that $\bar{\theta}_1$ is asymptotically normal:

ASSUMPTION 1. *Let $\Phi(t)$ be the cumulative distribution function of the standard normal random variable $Z \sim N(0, 1)$. There exists a sequence $\{\sigma_n\}_{n \in \mathbb{N}} \subset \mathbb{R}_+$ such that*

$$g(n) := \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\bar{\theta}_1 - \theta_*}{\sigma_n} \leq t\right) - \Phi(t) \right| \rightarrow 0 \text{ as } n \rightarrow \infty.$$

THEOREM 1. *Assume that $s > 0$ and $n = \lfloor N/k \rfloor$ are such that*

$$g(n) + \sqrt{\frac{s}{k}} < \frac{1}{2}. \tag{3}$$

Moreover, let assumption 1 be satisfied, and let $\zeta(n, s)$ solve the equation

$$\Phi(\zeta(n, s)) = \frac{1}{2} + g(n) + \sqrt{\frac{s}{k}}.$$

Then for any s satisfying (3),

$$\left| \widehat{\theta}^{(k)} - \theta_* \right| \leq \sigma_n \zeta(n, s)$$

with probability $\geq 1 - 4e^{-2s}$.

PROOF. See section 4.2.

The following Lemma yields a more “explicit” form of the bound:

LEMMA 1. Assume that $g(n) + \sqrt{\frac{s}{k}} \leq 0.33$. Then

$$\zeta(n, s) \leq \frac{\sqrt{2\pi}}{1 - 1.51 \left(g(n) + \sqrt{s/k} \right)^2} \left(g(n) + \sqrt{s/k} \right) < 3 \left(g(n) + \sqrt{s/k} \right).$$

PROOF. See section 4.8.

3.1.1. Example: new bounds for the median-of-means estimator.

The univariate mean estimation problem is pervasive in statistics, and serves as a building block of more advanced methods such as empirical risk minimization. Early works on robust mean estimation include Tukey’s “trimmed mean” (Tukey and Harris, 1946), as well as “winsorized mean” (Bickel et al., 1965); also see discussion in (Bubeck et al., 2013). These techniques often produce estimators with significant bias. A different approach based on M-estimation was suggested by O. Catoni (Catoni, 2012); Catoni’s estimator yields almost optimal constants, however, its construction requires additional information about the variance or the kurtosis of the underlying distribution; moreover, its computation is not easily parallelizable, therefore this technique cannot be easily employed in the distributed setting.

Here, we will focus on a fruitful idea that is commonly referred to as the “median-of-means” estimator that we formally defined in (1) above. Several refinements and extensions of this estimator to higher dimensions have been recently introduced by Minsker (2015); Hsu and Sabato (2013); Devroye et al. (2016); Joly et al. (2016); Lugosi and Mendelson (2017). Advantages of this method include the facts that it can be implemented in parallel and does not require prior knowledge of any information about parameters of the distribution (e.g., its variance).

The following result for the median-of-means estimator is the corollary of Theorem 1:

COROLLARY 1. Let X_1, \dots, X_N be a sequence of i.i.d. copies of a random variable X such that $\mathbb{E}X = \theta_*$, $\text{Var}(X) = \sigma^2$, $\mathbb{E}|X - \theta_*|^3 < \infty$, and define $c_n := 0.4748 \frac{\mathbb{E}|X - \theta_*|^3}{\sigma^3 \sqrt{n}}$.

Then for all $s > 0$ such that $c_n + \sqrt{\frac{s}{k}} \leq 0.33$, the estimator $\hat{\theta}^{(k)}$ defined in (1) satisfies

$$|\hat{\theta}^{(k)} - \theta_*| \leq \sigma \left(1.43 \frac{\mathbb{E}|X - \theta_*|^3 / \sigma^3}{n} + 3\sqrt{\frac{s}{kn}} \right)$$

with probability $\geq 1 - 4e^{-2s}$.

REMARK 1. Note that whenever $k \lesssim \sqrt{N}$ (so that $n \gtrsim \sqrt{N}$), the right-hand side of the inequality above is of “optimal” order $(kn)^{-1/2} \simeq N^{-1/2}$.

PROOF. It follows from the Berry-Essen Theorem (fact 1) that assumption 1 is satisfied with $\sigma_n = \frac{\sigma}{\sqrt{n}}$ and

$$g(n) = c_n := 0.4748 \frac{\mathbb{E}|X - \theta_*|^3}{\sigma^3 \sqrt{n}}.$$

Lemma 1 implies that

$$\zeta(n, s) \leq 3 \left(c_n + \sqrt{s/k} \right).$$

and the claim follows from Theorem 1.

3.1.2. Example: distributed Maximum Likelihood Estimation.

Let X_1, \dots, X_N be i.i.d. copies of a random vector $X \in \mathbb{R}^d$ with distribution P_{θ_*} , where $\theta_* \in \Theta \subseteq \mathbb{R}$. Assume that for each $\theta \in \Theta$, P_θ is absolutely continuous with respect to a σ -finite measure μ , and let $p_\theta = \frac{dP_\theta}{d\mu}$ be the corresponding density. In this section, we state sufficient conditions for assumption 1 to be satisfied when $\bar{\theta}_1, \dots, \bar{\theta}_k$ are the maximum likelihood estimators (van der Vaart, 1998) of θ_* . Conditions stated below were obtained by Pinelis (2016). All derivatives below (denoted by $'$) are taken with respect to θ , unless noted otherwise.

Assume that the log-likelihood function $\ell_x(\theta) = \log p_\theta(x)$ satisfies the following:

- (1) $[\theta_* - \delta, \theta_* + \delta] \subseteq \Theta$ for some $\delta > 0$;
- (2) “standard regularity conditions” that allow differentiation under the expectation: assume that $\mathbb{E}\ell'_X(\theta_*) = 0$, and that the Fisher information $\mathbb{E}\ell''_X(\theta_*)^2 = -\mathbb{E}\ell'''_X(\theta_*) := I(\theta_*)$ is finite;
- (3) $\mathbb{E}|\ell'_X(\theta_*)|^3 + \mathbb{E}|\ell''_X(\theta_*)|^3 < \infty$;
- (4) for μ -almost all x , $\ell_x(\theta)$ is three times differentiable for $\theta \in [\theta_* - \delta, \theta_* + \delta]$, and $\mathbb{E}\sup_{|\theta - \theta_*| \leq \delta} |\ell'''_X(\theta)|^3 < \infty$;
- (5) $\mathbb{P}(|\bar{\theta}_1 - \theta_*| \geq \delta) \leq c\gamma^n$ for some positive constants c and $\gamma \in [0, 1)$.

In turn, condition (5) above is implied by the following two inequalities (see Pinelis, 2016, section 6.2, for detailed discussion and examples):

- (a) $H^2(\theta, \theta_*) \geq 2 - \frac{2}{(1+c_0(\theta-\theta_*)^2)^\gamma}$, where $H(\theta_1, \theta_2) = \sqrt{\int_{\mathbb{R}^d} (\sqrt{p_{\theta_1}} - \sqrt{p_{\theta_2}})^2 d\mu}$ is the Hellinger distance, and c_0, γ are positive constants;
- (b) $I(\theta) \leq c_1 + c_2 |\theta|^\alpha$ for some positive constants c_1, c_2 and α and all $\theta \in \Theta$.

COROLLARY 2. *Assume that conditions (1)-(5) are satisfied. Then for all $s > 0$ such that $\frac{\mathfrak{C}}{\sqrt{n}} + c\gamma^n + \sqrt{\frac{s}{k}} \leq 0.33$,*

$$|\hat{\theta}^{(k)} - \theta_*| \leq \frac{3}{\sqrt{I(\theta_*)}} \left(\frac{\mathfrak{C}}{n} + \frac{c}{\sqrt{n}} \gamma^n + \sqrt{\frac{s}{kn}} \right)$$

with probability $\geq 1 - 4e^{-2s}$, where \mathfrak{C} is a positive constant that depends only on $\{P_\theta\}_{\theta \in [\theta_* - \delta, \theta_* + \delta]}$.

PROOF. It follows from results in (Pinelis, 2016) (in particular, equation (5.5)) that, whenever conditions (1)-(5) hold, assumption 1 is satisfied with $\sigma_n = (nI(\theta_*))^{-1/2}$, where $I(\theta_*)$ is the Fisher information, and $g(n) = \frac{\mathfrak{C}}{\sqrt{n}} + c\gamma^n$, where \mathfrak{C} is a constant that depends only on $\{P_\theta\}_{\theta \in [\theta_* - \delta, \theta_* + \delta]}$. Lemma 1 implies that

$$\zeta(n, s) \leq 3 \left(\frac{\mathfrak{C}}{\sqrt{n}} + c\gamma^n + \sqrt{s/k} \right),$$

and the claim follows from Theorem 1.

3.2. Connection to U-quantiles.

In this section, we discuss connections of proposed algorithms to U-quantiles and the assumption requiring the groups G_1, \dots, G_k to be disjoint. As before, assume that X_1, \dots, X_N is a collection of i.i.d. random variables with common distribution P , and let $\theta_* = \theta_*(P) \in \mathbb{R}$ be a real-valued parameter of interest. It is clear that the estimators produced by distributed algorithms considered above depend on the random partition of the sample. A natural way to avoid such dependence is to consider the U-quantile (in this case, the median)

$$\tilde{\theta}^{(k)} = \text{med} \left(\bar{\theta}_J, J \in \mathcal{A}_N^{(n)} \right),$$

where $\mathcal{A}_N^{(n)} := \{J : J \subseteq \{1, \dots, N\}, \text{card}(J) = n := \lfloor N/k \rfloor\}$ is a collection of all distinct subsets of $\{1, \dots, N\}$ of cardinality n , and $\bar{\theta}_J := \bar{\theta}(X_j, j \in J)$ is an estimator of θ_* based on $\{X_j, j \in J\}$. For instance, when $\text{card}(J) = 2$ and $\bar{\theta}_J = \frac{1}{\text{card}(J)} \sum_{j \in J} \frac{X_j}{2}$, $\tilde{\theta}^{(k)}$ is the well-known Hodges-Lehmann estimator of the location parameter, see (Hodges and Lehmann, 1963; Lehmann and D'Abrera, 2006); for a comprehensive study of U-quantiles, see (Arcones, 1996). The main result of this section is an analogue of Theorem 1 for the estimator $\tilde{\theta}^{(k)}$; it implies that theoretical guarantees for the performance of $\tilde{\theta}^{(k)}$ are at least as good as for the estimator $\hat{\theta}^{(k)}$:

THEOREM 2. *Assume that $s > 0$ and $n = \lfloor N/k \rfloor$ are such that*

$$g(n) + \sqrt{\frac{s}{k}} < \frac{1}{2}. \tag{4}$$

Moreover, let assumption 1 be satisfied, and let $\zeta(n, s)$ solve the equation

$$\Phi(\zeta(n, s)) = \frac{1}{2} + g(n) + \sqrt{\frac{s}{k}}.$$

Then for any s satisfying (4),

$$\left| \tilde{\theta}^{(k)} - \theta_* \right| \leq \sigma_n \zeta(n, s)$$

with probability $\geq 1 - 4e^{-2s}$.

PROOF. See section 4.3.

As before, a more explicit form of the bound immediately follows from Lemma 1. A drawback of the estimator $\tilde{\theta}^{(k)}$ is the fact that its exact computation requires evaluation of $\binom{n}{N}$ estimators $\bar{\theta}_J$ over subsamples $\left\{ \{X_j, j \in J\}, J \in \mathcal{A}_N^{(n)} \right\}$. For large N and n , such task becomes intractable. However, an approximate result can be obtained by choosing ℓ subsets J_1, \dots, J_ℓ from $\mathcal{A}_N^{(n)}$ uniformly at random, and setting $\tilde{\theta}_\ell^{(k)} := \text{med}(\bar{\theta}_{J_1}, \dots, \bar{\theta}_{J_\ell})$. Typically, the error $\left| \tilde{\theta}_\ell^{(k)} - \tilde{\theta}^{(k)} \right|$ is of order $\ell^{-1/2}$ with high probability over the random draw J_1, \dots, J_ℓ .

3.3. Estimation in higher dimensions.

In this section, we will assume that $\theta_* = \theta_*(P) \in \mathbb{R}^m$ is a vector-valued parameter of interest. As before, let $X_1, \dots, X_N \in \mathbb{R}^d$ be i.i.d. copies of X randomly partitioned into disjoint groups G_1, \dots, G_k of cardinality $n = \lfloor N/k \rfloor$ each. Let $\bar{\theta}_j := \bar{\theta}_j(G_j) \in \mathbb{R}^m$, $1 \leq j \leq k$ be a sequence of i.i.d. estimators of θ_* , and

$$\widehat{\theta}^{(k)} = \text{med}_g(\bar{\theta}_1, \dots, \bar{\theta}_k) := \underset{z \in \mathbb{R}^m}{\text{argmin}} \sum_{j=1}^k \|z - \bar{\theta}_j\|_2 \quad (5)$$

be the geometric (spatial) median of $\bar{\theta}_1, \dots, \bar{\theta}_k$.

Let $Z \in \mathbb{R}^m$ have multivariate normal distribution $N(0, \Sigma)$, and define $\Phi_\Sigma(A) := \mathbb{P}(Z \in A)$ for a Borel measurable set $A \subseteq \mathbb{R}^m$. Moreover, define \mathcal{S} to be the set of closed cones,

$$\mathcal{S}_m = \{C_u(t; b) = \{x \in \mathbb{R}^m : \langle x - b, u \rangle \geq t\|x - b\|_2\}, \|u\|_2 = 1, b \in \mathbb{R}^m, 0 \leq t \leq 1\}. \quad (6)$$

We will assume that $\bar{\theta}_1$ is “asymptotically normal on cones”:

ASSUMPTION 2. *There exists a sequence $\{\sigma_n\}_{n \in \mathbb{N}} \subset \mathbb{R}_+$ and a positive-definite matrix Σ such that $\|\Sigma\| \leq 1$ and*

$$g_{\mathcal{S}_m}(n) := \sup_{S \in \mathcal{S}_m} \left| \mathbb{P}\left(\frac{1}{\sigma_n} (\bar{\theta}_1 - \theta_*) \in S\right) - \Phi_\Sigma(S) \right| \rightarrow 0 \text{ as } n \rightarrow \infty.$$

THEOREM 3. *Let assumption 2 be satisfied. Then with probability $\geq 1 - e^{-2s}$,*

$$\tanh\left(\frac{1}{\sigma_n}\left\|\hat{\theta}^{(k)} - \theta_*\right\|_2\right) \leq 26.8\left\|\Sigma^{-1/2}\right\|\left(\frac{C_1(m)}{\sqrt{k}} + C_2(m)\left(\sqrt{\frac{s}{4k}} + g_{S_m}(n)\right)\right), \quad (7)$$

where

$$C_1(m) = 6\sqrt{\log 4e^{5/2}(m+4)}\sqrt{m+2\sqrt{(m-1)\ln 4}}$$

$$\text{and } C_2(m) = \sqrt{m+2\sqrt{(m-1)\ln 4}}.$$

PROOF. See section 4.4.

REMARK 2. *It follows from Lemma 5 that whenever the right-hand side of the inequality (7) is bounded by $1/2$, $\tanh\left(\frac{1}{\sigma_n}\left\|\hat{\theta}^{(k)} - \theta_*\right\|_2\right) \geq \frac{0.83}{\sigma_n}\left\|\hat{\theta}^{(k)} - \theta_*\right\|_2$, which leads to a more explicit bound for $\left\|\hat{\theta}^{(k)} - \theta_*\right\|_2$.*

3.3.1. Example: multivariate median-of-means estimator.

We consider the special case of Theorem 3 when $\theta_* = \mathbb{E}X$ is the mean of X and $\bar{\theta}_j(X) := \frac{1}{|G_j|} \sum_{X_i \in G_j} X_i$ is the sample mean evaluated over the subsample G_j . The problem of finding a mean estimator that admits sub-Gaussian concentration around $\mathbb{E}X$ under weak moment assumptions on the underlying distribution has recently been investigated in several works. For instance, Joly et al. (2016) construct an estimator that admits “almost optimal” behavior under the assumption that the entries of X possess 4 moments. Recently, Lugosi and Mendelson (2017) proposed a new estimator that attains optimal bounds and requires existence of only 2 moments. More specifically, the aforementioned paper shows that, for any $0 < \delta < 1$, there is an estimator $\hat{\theta}_{(\delta)}$ such that with probability $\geq 1 - \delta$,

$$\left\|\hat{\theta}_{(\delta)} - \theta_*\right\|_2 \leq C\left(\sqrt{\frac{\text{tr}(\tilde{\Sigma})}{N}} + \sqrt{\frac{\lambda_{\max}(\tilde{\Sigma}) \log(2/\delta)}{N}}\right),$$

where $C > 0$ is a numerical constant, $\tilde{\Sigma}$ is the covariance matrix of X , $\text{tr}(\tilde{\Sigma})$ is its trace and $\lambda_{\max}(\tilde{\Sigma})$ - its largest eigenvalue. However, both of these estimators are difficult to compute, and their construction depends on the desired estimation confidence level $1 - \delta$.

Minsker (2015) provides the analysis of the multivariate median-of-means estimator (5) that can be evaluated numerically, however, resulting bounds are suboptimal (more precisely, it was shown that, for $k = \lfloor \log(1/\delta) \rfloor$, $\left\|\hat{\theta}^{(k)} - \theta_*\right\|_2 \leq C\sqrt{\frac{\text{tr}(\tilde{\Sigma}) \log(1/\delta)}{N}}$ with probability $\geq 1 - \delta$).

Results presented in this section provide new insights into the properties of the multivariate median-of-means estimator. In particular, we show that its performance is robust with respect to the choice of k , the number of subgroups, whenever the entries of X have finite third moments. In other words, deviation bounds for the estimator (5) hold *simultaneously* over a wide range of confidence levels; moreover, resulting bounds are optimal with respect to the sample size N whenever $k \lesssim \sqrt{N}$.

COROLLARY 3. Let X_1, \dots, X_N be a sequence of i.i.d. copies of a random vector $X \in \mathbb{R}^d$ such that $\mathbb{E}X = \theta_*$, $\mathbb{E}[(X - \theta_*)(X - \theta_*)^T] = \tilde{\Sigma}$, and $\mathbb{E}\|X - \theta_*\|_2^3 < \infty$. Define

$$\hat{\theta}^{(k)} = \text{med}_g(\bar{\theta}_1, \dots, \bar{\theta}_k).$$

Assume that $s > 0$ and $k \leq N/2$ are such that

$$\text{cond}(\tilde{\Sigma}^{1/2}) \left(\frac{C_1(d)}{\sqrt{k}} + C_2(d) \left(\sqrt{\frac{s}{4k}} + \frac{400d^{1/4} \mathbb{E} \left\| \tilde{\Sigma}^{-1/2}(X - \theta_*) \right\|_2^3}{\sqrt{n}} \right) \right) \leq 0.037.$$

Then

$$\begin{aligned} \left\| \hat{\theta}^{(k)} - \theta_* \right\|_2 &\leq 32.4 \|\tilde{\Sigma}^{1/2}\| \text{cond}(\tilde{\Sigma}^{1/2}) \\ &\quad \times \left(\frac{C_1(d)}{\sqrt{kn}} + C_2(d) \left(\sqrt{\frac{s}{4kn}} + \frac{400d^{1/4} \mathbb{E} \left\| \tilde{\Sigma}^{-1/2}(X - \theta_*) \right\|_2^3}{n} \right) \right) \end{aligned}$$

with probability $\geq 1 - e^{-2s}$, where $C_1(d)$ and $C_2(d)$ are the same as in Theorem 3.

PROOF. It follows from the multivariate Berry-Esseen bound (fact 2) that assumption 2 is satisfied with $\sigma_n = \sqrt{\frac{\|\tilde{\Sigma}\|}{n}}$, $\Sigma = \frac{\tilde{\Sigma}}{\|\tilde{\Sigma}\|}$ and $g_{S_d}(n) = \frac{400d^{1/4} \mathbb{E} \left\| \tilde{\Sigma}^{-1/2} X \right\|_2^3}{\sqrt{n}}$. Noting that $\|\Sigma^{-1/2}\| = \|\tilde{\Sigma}^{1/2}\| \|\tilde{\Sigma}^{-1/2}\| = \text{cond}(\tilde{\Sigma}^{1/2})$, it is easy to deduce the bound from (7) and remark 2.

REMARK 3. Note that, similarly to the case $d = 1$, whenever $k \lesssim \sqrt{N}$ (hence, $n \gtrsim \sqrt{N}$), the bound of Corollary 3 is of “optimal” order $(kn)^{-1/2} \simeq N^{-1/2}$ with respect to the sample size N . However, dependence of the bound on the dimension factor d is suboptimal. While improved (with respect to d) bounds exist for certain values of the confidence parameter s (e.g., $s \simeq k$, see (Minsker, 2015)), it is unclear if the median-of-means estimator satisfies deviation inequalities with optimal dependence on the dimension in general.

We illustrate results of this section with a numerical simulation that compares performance of the median-of-means estimator with the usual sample mean, see figure 2 below.

4. Proofs

In this section, we outline the proofs of the main results.

4.1. Preliminaries

We recall several facts that are used in the proofs below. The following bound has been established by A. Berry (Berry, 1941) and C.-G. Esseen (Esseen, 1942). A version with an explicit constant given below is due to Shevtsova (2011).

FACT 1 (BERRY-ESSEEN BOUND). Assume that Y_1, \dots, Y_n is a sequence of i.i.d. copies of a random variable Y with mean μ , variance σ^2 and such that $\mathbb{E}|Y|^3 < \infty$. Then

$$\sup_{s \in \mathbb{R}} \left| \mathbb{P} \left(\sqrt{n} \frac{\bar{Y} - \mu}{\sigma} \leq s \right) - \Phi(s) \right| \leq 0.4748 \frac{\mathbb{E}|Y|^3}{\sigma^3 \sqrt{n}},$$

where $\bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j$ and $\Phi(s)$ is the cumulative distribution function of the standard normal random variable.

Everywhere below, Φ_Σ stands for the distribution of the normal vector with mean 0 and covariance matrix Σ . The following multivariate version of the Berry-Esseen Theorem for convex sets has been established by [Bentkus \(2003\)](#).

FACT 2 (MULTIVARIATE BERRY-ESSEEN BOUND). Assume that Y_1, \dots, Y_n is a sequence of i.i.d. copies of a random vector $Y \in \mathbb{R}^d$ with mean μ , covariance matrix $\Sigma \succ 0$ and such that $\mathbb{E}\|Y\|_2^3 < \infty$. Let Z have normal distribution $N(0, \Sigma)$, and \mathcal{A} be the class of all convex subsets of \mathbb{R}^d . Then

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}(\sqrt{n}(\bar{Y} - \mu) \in A) - \Phi_\Sigma(A) \right| \leq \frac{400d^{1/4} \mathbb{E}\|\Sigma^{-1/2}Y\|_2^3}{\sqrt{n}},$$

where $\bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j$.

Next, we recall two useful concentration inequalities.

FACT 3 (BOUNDED DIFFERENCE INEQUALITY). Let X_1, \dots, X_n be i.i.d. random variables, and assume that $Z = g(X_1, \dots, X_n)$, where g is such that for all $j = 1, \dots, n$ and all $x_1, x_2, \dots, x_j, x'_j, \dots, x_n$,

$$\left| g(x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_n) - g(x_1, \dots, x_{j-1}, x'_j, x_{j+1}, \dots, x_n) \right| \leq c_j.$$

Then

$$\mathbb{P}(Z - \mathbb{E}Z \geq t) \leq \exp \left\{ -\frac{2t^2}{\sum_{j=1}^n c_j^2} \right\}$$

and

$$\mathbb{P}(Z - \mathbb{E}Z \leq -t) \leq \exp \left\{ -\frac{2t^2}{\sum_{j=1}^n c_j^2} \right\}.$$

Next, we recall the definition of a U-statistic. Let $h : \mathbb{R}^n \mapsto \mathbb{R}$ be a measurable function of n variables, and

$$\mathcal{A}_N^{(n)} := \{J : J \subseteq \{1, \dots, N\}, \text{card}(J) = n\}.$$

A U-statistic of order n with kernel h based on the i.i.d. sample X_1, \dots, X_N is defined as ([Hoeffding, 1948](#))

$$U_N(h) = \frac{1}{\binom{N}{n}} \sum_{J \in \mathcal{A}_N^{(n)}} h(X_j, j \in J).$$

Clearly, $\mathbb{E}U_N(h) = \mathbb{E}h(X_1, \dots, X_n)$, moreover, $U_N(h)$ has the smallest variance among all unbiased estimators.

FACT 4 (CONCENTRATION INEQUALITY FOR U-STATISTICS, ([Hoeffding, 1963](#))).

Assume that the kernel h satisfies $|h(x_1, \dots, x_n)| \leq M$ for all x_1, \dots, x_n . Then for all $s > 0$,

$$\mathbb{P}(|U_N(h) - \mathbb{E}U_N(h)| \geq s) \leq 2 \exp \left\{ -\frac{2 \lfloor N/n \rfloor t^2}{M^2} \right\}.$$

Given a metric space (T, ρ) , the covering number $N(T, \rho, \varepsilon)$ is defined as the smallest $N \in \mathbb{N}$ such that there exists a subset $F \subseteq T$ of cardinality N with the property that for all $z \in T$, $\rho(z, F) \leq \varepsilon$. When metric ρ is clear from the context, we will simply write $N(T, \varepsilon)$.

Let $\{Y(t), t \in T\}$ be a stochastic process indexed by T . We will say that it has sub-Gaussian increments with respect to metric ρ if for all $t_1, t_2 \in T$ and $s > 0$,

$$\mathbb{P}(|Y_{t_1} - Y_{t_2}| \geq s\rho(t_1, t_2)) \leq 2e^{-s^2/2}.$$

FACT 5 (DUDLEY'S ENTROPY BOUND). Let $\{Y(t), t \in T\}$ be a centered stochastic process with sub-Gaussian increments. Then the following inequality holds:

$$\mathbb{E} \sup_{t \in T} Y(t) \leq 12 \int_0^{D(T)} \sqrt{\log N(T, \rho, \varepsilon)} d\varepsilon, \quad (8)$$

where $D(T)$ is the diameter of the space T with respect to ρ .

PROOF. See ([Talagrand, 2005](#)).

Finally, we recall two useful facts related to Vapnik-Chervonenkis (VC) combinatorics (see [van der Vaart and Wellner, 1996](#), for the definition of VC dimension and related theory). Let \mathcal{F} be a finite-dimensional vector space of real functions on S .

FACT 6. Let $\mathcal{C} = \{\{f \geq 0\} : f \in \mathcal{F}\}$ and $\mathcal{C}_+ = \{\{f > 0\} : f \in \mathcal{F}\}$. Then

$$\text{VC}(\mathcal{C}) = \text{VC}(\mathcal{C}_+) = \dim(\mathcal{F}).$$

PROOF. See Proposition 3.6.6 in ([Giné and Nickl, 2015](#)).

FACT 7. Let \mathcal{C} be a class of sets of VC-dimension V . Then, for any probability measure Q ,

$$N(\mathcal{C}, L_2(Q), \varepsilon) \leq e(V+1)(4e)^V \left(\frac{1}{\varepsilon^2} \right)^V \quad (9)$$

for all $0 < \varepsilon \leq 1$;

PROOF. This bound follows from results of R. Dudley ([Dudley, 1978](#)) and D. Haussler ([Haussler, 1995](#)). The bound with explicit constants as stated above is given in ([Pollard, 2000](#)).

4.2. Proof of Theorem 1.

Observe that

$$\left| \widehat{\theta}^{(k)} - \theta_* \right| = \sigma_n \left| \text{med} \left(\frac{\bar{\theta}_1 - \theta_*}{\sigma_n}, \dots, \frac{\bar{\theta}_k - \theta_*}{\sigma_n} \right) \right|.$$

Let $\Phi^{(n)}(\cdot)$ be the distribution function of $\frac{\bar{\theta}_1 - \theta_*}{\sigma_n}$ and $\Phi_k^{(n)}(\cdot)$ - the empirical distribution function corresponding to the sample $W_1 = \frac{\bar{\theta}_1 - \theta_*}{\sigma_n}, \dots, W_k = \frac{\bar{\theta}_k - \theta_*}{\sigma_n}$. Suppose that $z \in \mathbb{R}$ is fixed, and note that $\Phi_k^{(n)}(z)$ is a function of the random variables W_1, \dots, W_k , and $\Phi_k^{(n)}(z) = \mathbb{E}\Phi^{(n)}(z)$. Moreover, the hypothesis of the bounded difference inequality (fact 3) is satisfied with $c_j = 1/k$ for $j = 1, \dots, k$, and therefore it implies that

$$\left| \Phi_k^{(n)}(z) - \Phi^{(n)}(z) \right| \leq \sqrt{\frac{s}{k}} \quad (10)$$

on the draw of W_1, \dots, W_k with probability $\geq 1 - 2e^{-2s}$.

Let $z_1 \geq z_2$ be such that $\Phi^{(n)}(z_1) \geq \frac{1}{2} + \sqrt{\frac{s}{k}}$ and $\Phi^{(n)}(z_2) \leq \frac{1}{2} - \sqrt{\frac{s}{k}}$. Applying (10) for $z = z_1$ and $z = z_2$ together with the union bound, we see that for $j = 1, 2$,

$$\left| \Phi_k^{(n)}(z_j) - \Phi^{(n)}(z_j) \right| \leq \sqrt{\frac{s}{k}}$$

on an event \mathcal{E} of probability $\geq 1 - 4e^{-2s}$. It follows that on \mathcal{E} , $\Phi_k^{(n)}(z_1) \geq 1/2$ and $1 - \Phi_k^{(n)}(z_2) \geq 1/2$ simultaneously, hence

$$\text{med}(W_1, \dots, W_k) \in [z_2, z_1] \quad (11)$$

by the definition of the median. It remains to estimate z_1 and z_2 . Assumption 1 implies that

$$\Phi^{(n)}(z_1) \geq \Phi(z_1) - |\Phi^{(n)}(z_1) - \Phi(z_1)| \geq \Phi(z_1) - g(n).$$

Hence, it suffices to find z_1 such that $\Phi(z_1) \geq \frac{1}{2} + g(n) + \sqrt{\frac{s}{k}}$. If $\zeta(n, s)$ be the solution of the equation

$$\Phi(\zeta(n, s)) = \frac{1}{2} + g(n) + \sqrt{\frac{s}{k}},$$

then clearly any $z_1 \geq \zeta(n, s)$ satisfies the requirements (note that $\zeta(n, s)$ always exists since $g(n) + \sqrt{\frac{s}{k}} < \frac{1}{2}$ by assumption). Similarly,

$$\Phi^{(n)}(z_2) \leq \Phi(z_2) + |\Phi^{(n)}(z_2) - \Phi(z_2)| \leq \Phi(z_2) + g(n)$$

by assumption 1, hence it is sufficient to choose z_2 such that $z_2 \leq \zeta_2(n, s)$, where $\zeta_2(n, s)$ satisfies $\Phi(\zeta_2(n, s)) = \frac{1}{2} - g(n) - \sqrt{\frac{s}{k}}$. Noting that $\zeta_2(n, s) = -\zeta(n, s)$ and recalling (11), we conclude that

$$\left| \widehat{\theta}^{(k)} - \theta_* \right| \leq \sigma_n \zeta(n, s)$$

with probability $\geq 1 - 4e^{-2s}$.

4.3. Proof of Theorem 2.

The argument is similar to the proof of Theorem 1. Let $\Phi^{(n)}(\cdot)$ be the distribution function of $\frac{\bar{\theta}_1 - \theta_*}{\sigma_n}$ and $\Phi_{\binom{n}{N}}^{(n)}(\cdot)$ - the empirical distribution function corresponding to the sample $\left\{W_J = \frac{\bar{\theta}_J - \theta_*}{\sigma_n}, J \in \mathcal{A}_N^{(n)}\right\}$ of size $\binom{n}{N}$.

Suppose that $z \in \mathbb{R}$ is fixed, and note that $\Phi_{\binom{n}{N}}^{(n)}(z)$ is a U-statistic with mean $\Phi^{(n)}(z)$. We will apply the concentration inequality for U-statistics (fact 4) with $M = 1$ to get that

$$\left| \Phi_{\binom{n}{N}}^{(n)}(z) - \Phi^{(n)}(z) \right| \leq \sqrt{\frac{s}{\lfloor N/n \rfloor}} \leq \sqrt{\frac{s}{k}} \quad (12)$$

with probability $\geq 1 - 2e^{-2s}$; here, we also used the fact that $n = \lfloor N/k \rfloor$.

Let $z_1 \geq z_2$ be such that $\Phi^{(n)}(z_1) \geq \frac{1}{2} + \sqrt{\frac{s}{k}}$ and $\Phi^{(n)}(z_2) \leq \frac{1}{2} - \sqrt{\frac{s}{k}}$. Applying (12) for $z = z_1$ and $z = z_2$ together with the union bound, we see that for $j = 1, 2$,

$$\left| \Phi_{\binom{n}{N}}^{(n)}(z_j) - \Phi^{(n)}(z_j) \right| \leq \sqrt{\frac{s}{k}}$$

on an event \mathcal{E} of probability $\geq 1 - 4e^{-2s}$. It follows that on \mathcal{E} , $\text{med}\left(W_J, J \in \mathcal{A}_N^{(n)}\right) \in [z_2, z_1]$. The rest of the proof repeats the argument of section 4.2.

4.4. Proof of Theorem 3.

By the definition of the geometric median,

$$\widehat{\theta}^{(k)} = \underset{z \in \mathbb{R}^m}{\operatorname{argmin}} \sum_{j=1}^k \|z - \bar{\theta}_j\|_2,$$

hence

$$\frac{1}{\sigma_n} \left(\widehat{\theta}^{(k)} - \theta_* \right) = \underset{z \in \mathbb{R}^m}{\operatorname{argmin}} \sum_{j=1}^k \left\| z - \frac{1}{\sigma_n} (\bar{\theta}_j - \theta_*) \right\|_2. \quad (13)$$

Set $F_k(z) := \sum_{j=1}^k \left\| z - \frac{1}{\sigma_n} (\bar{\theta}_j - \theta_*) \right\|_2$. Then (13) is equivalent to

$$\widehat{\mu}^{(k)} := \frac{1}{\sigma_n} \left(\widehat{\theta}^{(k)} - \theta_* \right) = \underset{z \in \mathbb{R}^m}{\operatorname{argmin}} F(z).$$

Denote by $\Phi^{(n)}$ the distribution of $\frac{1}{\sigma_n} (\bar{\theta}_1 - \mu)$, and by $\Phi_k^{(n)}$ - the empirical distribution corresponding to the sample

$$W_1 = \frac{1}{\sigma_n} (\bar{\theta}_1 - \mu), \dots, W_k = \frac{1}{\sigma_n} (\bar{\theta}_k - \mu).$$

Let $DF_k(\hat{\mu}^{(k)}; u) := \lim_{t \searrow 0} \frac{F_k(\hat{\mu}^{(k)} + tu) - F_k(\hat{\mu}^{(k)})}{t}$ be the directional derivative of F_k at point $\hat{\mu}^{(k)}$ in direction u . Clearly, $DF_k(\hat{\mu}^{(k)}; u) \geq 0$ for any u such that $\|u\|_2 = 1$. On the other hand, it is easy to check that $DF_k(\hat{\mu}^{(k)}; u) = \Phi_k^{(n)} f_{u, \hat{\mu}^{(k)}}$, where

$$f_{u,b}(x) = \begin{cases} \left\langle \frac{x-b}{\|x-b\|_2}, u \right\rangle, & x \neq b, \\ 1, & x = b. \end{cases}$$

Let \mathcal{S}_m be the set of closed cones defined in (6), and note that for any unit vector $u \in \mathbb{R}^m$ and $t \in [0, 1]$,

$$\{x \in \mathbb{R}^m : f_{u, \hat{\mu}^{(k)}}(x) \geq t\} = C_u(t; \hat{\mu}^{(k)}). \quad (14)$$

Next, observe that

$$0 \leq DF(\hat{\mu}^{(k)}; u) = (\Phi_k^{(n)} - \Phi^{(n)})f_{u, \hat{\mu}^{(k)}} + (\Phi^{(n)} - \Phi_\Sigma)f_{u, \hat{\mu}^{(k)}} + \Phi_\Sigma f_{u, \hat{\mu}^{(k)}}. \quad (15)$$

We will assume that u is *chosen* such that $\Phi_\Sigma f_{u, \hat{\mu}^{(k)}} \leq 0$ (if not, simply replace u by $-u$). Then (15) implies that

$$\Phi_\Sigma f_{-u, \hat{\mu}^{(k)}} \leq \left| (\Phi_k^{(n)} - \Phi^{(n)})f_{u, \hat{\mu}^{(k)}} \right| + \left| (\Phi^{(n)} - \Phi_\Sigma)f_{u, \hat{\mu}^{(k)}} \right|. \quad (16)$$

It remains to estimate the left-hand side of inequality (16) from below and its right-hand side from above. We start by finding an upper bound (proved in section 4.5) for $\left| (\Phi^{(n)} - \Phi_\Sigma)f_{u, \hat{\mu}^{(k)}} \right|$.

LEMMA 2. *The following bound holds:*

$$\left| (\Phi^{(n)} - \Phi_\Sigma)f_{u, \hat{\mu}^{(k)}} \right| \leq 2g_{\mathcal{S}_m}(n),$$

where $g_{\mathcal{S}_m}(n)$ was defined in assumption 2.

The next Lemma (proved in section 4.6) provides an upper bound for $\left| (\Phi_k^{(n)} - \Phi^{(n)})f_{u, \hat{\mu}^{(k)}} \right|$.

LEMMA 3. *With probability $\geq 1 - e^{-2s}$,*

$$\left| (\Phi_k^{(n)} - \Phi^{(n)})f_{u, \hat{\mu}^{(k)}} \right| \leq \frac{12(m+4)}{\sqrt{k}} \sqrt{\log 4e^{5/2}} + \sqrt{\frac{s}{k}}.$$

Finally, it remains to estimate $\Phi_\Sigma f_{-u, \hat{\mu}^{(k)}}$ from below. The following inequality (proved in section 4.7) holds:

LEMMA 4. *Set $u = -\frac{\Sigma^{-1}\hat{\mu}^{(k)}}{\|\Sigma^{-1}\hat{\mu}^{(k)}\|_2}$. Then*

$$\Phi_\Sigma f_{-u, \hat{\mu}^{(k)}} \geq \frac{0.15}{2 \|\Sigma^{-1/2}\| \sqrt{m + 2\sqrt{(m-1)\ln 4}}} \tanh\left(\left\|\hat{\mu}^{(k)}\right\|_2\right),$$

where $\tanh(\cdot)$ is the hyperbolic tangent defined as $\tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}}$.

It therefore follows from Lemmas 2, 3 and 4 that with probability exceeding $1 - e^{-2s}$,

$$\frac{0.15}{2 \|\Sigma^{-1/2}\| \sqrt{m + 2\sqrt{(m-1)\ln 4}}} \tanh\left(\left\|\hat{\mu}^{(k)}\right\|_2\right) \leq \frac{12(m+4)}{\sqrt{k}} \sqrt{\log 4e^{5/2}} + \sqrt{\frac{s}{k}} + 2g_S(n),$$

which implies the bound of Theorem 3.

4.5. Proof of Lemma 2.

Recall that for any non-negative function $f : \mathbb{R}^m \mapsto \mathbb{R}_+$ and a signed measure Q ,

$$\int_{\mathbb{R}^m} f(x) dQ = \int_0^\infty Q(x : f(x) \geq t) dt. \quad (17)$$

Hence

$$(\Phi^{(n)} - \Phi_\Sigma) f_{u, \hat{\mu}^{(k)}} = (\Phi^{(n)} - \Phi_\Sigma) \max(f_{u, \hat{\mu}^{(k)}}, 0) - (\Phi^{(n)} - \Phi_\Sigma) \max(f_{-u, \hat{\mu}^{(k)}}(x), 0),$$

where we used the identity $-f_{u, \hat{\mu}^{(k)}} = f_{-u, \hat{\mu}^{(k)}}$. Next, it follows from (14) that

$$\begin{aligned} \left| (\Phi^{(n)} - \Phi_\Sigma) \max(f_{u, \hat{\mu}^{(k)}}, 0) \right| &= \left| \int_0^1 (\Phi^{(n)} - \Phi_\Sigma)(x : f_{u, \hat{\mu}^{(k)}} \geq t) dt \right| \\ &\leq \max_{0 \leq t \leq 1} \left| (\Phi^{(n)} - \Phi_\Sigma) C_u(t; \hat{\mu}^{(k)}) \right| \leq g_{S_m}(n) \end{aligned}$$

by assumption 2. It implies that $|(\Phi^{(n)} - \Phi_\Sigma) f_{u, \hat{\mu}^{(k)}}| \leq 2g_{S_m}(n)$, as claimed.

4.6. Proof of Lemma 3.

Using (17) and proceeding as in the proof of Lemma 2, we obtain that

$$\left| (\Phi_k^{(n)} - \Phi^{(n)}) f_{u, \hat{\mu}^{(k)}} \right| \leq \max_{0 \leq t \leq 1} \left| (\Phi_k^{(n)} - \Phi^{(n)}) C_u(t; \hat{\mu}^{(k)}) \right| \leq \sup_{A \in \mathcal{S}_m} \left| \Phi_k^{(n)}(A) - \Phi^{(n)}(A) \right|.$$

It follows from the bounded difference inequality (fact 3) that for all $s > 0$,

$$\mathbb{P} \left(\sup_{A \in \mathcal{S}_m} \left| \Phi_k^{(n)}(A) - \Phi^{(n)}(A) \right| - \mathbb{E} \sup_{A \in \mathcal{S}_m} \left| \Phi_k^{(n)}(A) - \Phi^{(n)}(A) \right| \geq \sqrt{\frac{s}{k}} \right) \leq e^{-2s},$$

hence it is enough to control $\mathbb{E} \sup_{A \in \mathcal{S}_m} \left| \Phi_k^{(n)}(A) - \Phi^{(n)}(A) \right|$. To this end, we will estimate the covering numbers of the class of cones \mathcal{S} and use Dudley's integral bound (fact 5).

Given a vector $\mathbf{x} \in \mathbb{R}^m$, let x_1, \dots, x_m be its coordinates with respect to the standard Euclidean basis. Note that

$$\langle \mathbf{x} - b, u \rangle \geq t \|\mathbf{x} - b\|_2 \iff \langle \mathbf{x} - b, u \rangle \geq 0 \text{ and } \langle \mathbf{x} - b, u \rangle^2 \geq t^2 \|\mathbf{x} - b\|_2^2,$$

which is equivalent to $\sum_{i,j=1}^m \alpha_i \alpha_{i,j} x_i x_j + \sum_{j=1}^m \beta_j x_j + \gamma \geq 0$ and $\langle x - b, u \rangle \geq 0$, where $\alpha_{i,j}$, β_j , $i, j = 1, \dots, m$, and γ are functions of t , b_j and u_j , $j = 1, \dots, m$. In particular,

every element of $A \in \mathcal{S}_m$ is the intersection of a half-space $\{\mathbf{x} : \langle \mathbf{x} - b, u \rangle \geq 0\}$ and a set $\{\mathbf{x} : f(\mathbf{x}) \geq 0\}$, where f is a polynomial of degree 2 in m variables. The dimension of the space $V_{2,m}$ of polynomials of degree at most 2 is $\dim(V_{2,m}) = \binom{m+2}{2}$, hence the Vapnik-Chernonenkis dimension of the collection of sets $\mathcal{S}_{V_{2,m}} = \left\{ \{x : f(x) \geq 0\}, f \in V_{2,m} \right\}$ is $\tilde{m} := \binom{m+2}{2}$ by fact 6. It follows from fact 7 that for any probability measure Q ,

$$N(\mathcal{S}_{V_{2,m}}, L_2(Q), \varepsilon) \leq e(\tilde{m} + 1)(4e)^{\tilde{m}} \left(\frac{1}{\varepsilon^2} \right)^{\tilde{m}} \quad (18)$$

for all $0 < \varepsilon \leq 1$. It is also well known that (and can be deduced from the similar reasoning) that the VC-dimension of a collection \mathcal{S}_L of halfspaces of \mathbb{R}^m is $m + 1$, hence

$$N(\mathcal{S}_L, L_2(Q), \varepsilon) \leq e(m + 2)(4e)^{m+1} \left(\frac{1}{\varepsilon^2} \right)^{m+1}.$$

Given two collections of sets $\mathcal{C}_1, \mathcal{C}_2$, let $A_1^{(1)}, \dots, A_{N(\mathcal{C}_1, L_2(Q), \varepsilon)}^{(1)}$ and $A_1^{(2)}, \dots, A_{N(\mathcal{C}_2, L_2(Q), \varepsilon)}^{(2)}$ be the $L_2(Q)$ ε -nets of smallest cardinality for the classes of functions $\{I_A : A \in \mathcal{C}_1\}$ and $\{I_A : A \in \mathcal{C}_2\}$ respectively. Let $A' \in \mathcal{C}_1$, $A'' \in \mathcal{C}_2$, and assume without loss of generality that $\|A' - A_1^{(1)}\|_{L_2(Q)} \leq \varepsilon$ and $\|A'' - A_1^{(2)}\|_{L_2(Q)} \leq \varepsilon$. Then

$$\left\| I_{A'} I_{A''} - I_{A_1^{(1)}} I_{A_1^{(2)}} \right\|_{L_2(Q)} \leq 2\varepsilon,$$

which implies that the covering number of the class $\mathcal{D} = \{I_{A_1} I_{A_2}, A_1 \in \mathcal{C}_1, A_2 \in \mathcal{C}_2\}$ corresponding to intersections of elements of \mathcal{C}_1 and \mathcal{C}_2 satisfies

$$N(\mathcal{D}, L_2(Q), \varepsilon) \leq N(\mathcal{C}_1, L_2(Q), \varepsilon/2) N(\mathcal{C}_2, L_2(Q), \varepsilon/2).$$

In particular, the metric entropy of the class of cones \mathcal{S}_m can be bounded as

$$\log N(\mathcal{S}_m, L_2(Q), \varepsilon) \leq 2 \left(\binom{m+2}{2} + m + 1 \right) \log \frac{4e^{3/2}}{\varepsilon}$$

uniformly over all probability measures Q , hence fact 5 implies that

$$\begin{aligned} \mathbb{E} \sup_{A \in \mathcal{S}_m} \left| \Phi_k^{(n)}(A) - \Phi^{(n)}(A) \right| &\leq \frac{12}{\sqrt{k}} \mathbb{E} \left[\int_0^1 \sqrt{\log N(\mathcal{S}_m, L_2(\Phi_k^{(n)}), \varepsilon)} d\varepsilon \right] \\ &\leq \frac{12}{\sqrt{k}} \mathbb{E} \left[\sqrt{\int_0^1 \log N(\mathcal{S}_m, L_2(\Phi_k^{(n)}), \varepsilon) d\varepsilon} \right] \leq \frac{12(m+4)}{\sqrt{k}} \sqrt{\log 4e^{5/2}}. \end{aligned}$$

4.7. Proof of Lemma 4.

Making the change of variables $x = \Sigma^{1/2}z$, we obtain

$$\begin{aligned} \int_{\mathbb{R}^m} \left\langle \frac{x - \hat{\mu}^{(k)}}{\|x - \hat{\mu}^{(k)}\|_2}, u \right\rangle d\Phi_\Sigma(x) &= \int_{\mathbb{R}^m} \left\langle \frac{\Sigma^{1/2}(z - \Sigma^{-1/2}\hat{\mu}^{(k)})}{\|\Sigma^{1/2}(z - \Sigma^{-1/2}\hat{\mu}^{(k)})\|_2}, u \right\rangle d\Phi(z) \\ &\geq \left\| \Sigma^{1/2}u \right\|_2 \int_{\mathbb{R}^m} \left\langle \frac{z - \Sigma^{-1/2}\hat{\mu}^{(k)}}{\|z - \Sigma^{-1/2}\hat{\mu}^{(k)}\|_2}, \tilde{u} \right\rangle d\Phi(z), \end{aligned}$$

where $\tilde{u} = \frac{\Sigma^{1/2}u}{\|\Sigma^{1/2}u\|_2}$. Let $\kappa := \|\Sigma^{-1/2}\hat{\mu}^{(k)}\|_2$, and note that $\kappa \geq \|\hat{\mu}^{(k)}\|_2$ since $\|\Sigma\| \leq 1$ by assumption. Let V be any orthogonal transformation that maps $\Sigma^{-1/2}\hat{\mu}^{(k)}$ to κe_1 (here, e_1, \dots, e_m is the standard Euclidean basis of \mathbb{R}^m). Then, letting $y = V(z - \Sigma^{-1/2}\hat{\mu}^{(k)})$, we observe that

$$\int_{\mathbb{R}^m} \left\langle \frac{x - \hat{\mu}^{(k)}}{\|x - \hat{\mu}^{(k)}\|_2}, u \right\rangle d\Phi_{\Sigma}(x) \geq \|\Sigma^{1/2}u\|_2 \int_{\mathbb{R}^m} \left\langle \frac{y}{\|y\|_2}, V\tilde{u} \right\rangle d\Phi(y + \kappa e_1).$$

Setting $u = -\frac{\Sigma^{-1}\hat{\mu}^{(k)}}{\|\Sigma^{-1}\hat{\mu}^{(k)}\|_2}$, we obtain from the last inequality that

$$\int_{\mathbb{R}^m} \left\langle \frac{x - \hat{\mu}^{(k)}}{\|x - \hat{\mu}^{(k)}\|_2}, u \right\rangle d\Phi_{\Sigma}(x) \geq \frac{1}{\|\Sigma^{-1/2}\|} \int_{\mathbb{R}^m} \left\langle \frac{y}{\|y\|_2}, -e_1 \right\rangle d\Phi(y + \kappa e_1).$$

Set $y = (-t, z)$, where $t \in \mathbb{R}$ and $z \in \mathbb{R}^{m-1}$. We will also let ϕ_k denote the density (with respect to Lebesgue measure) of the standard normal distribution on \mathbb{R}^k . Then

$$\int_{\mathbb{R}^m} \left\langle \frac{y}{\|y\|_2}, -e_1 \right\rangle d\Phi(y + \kappa e_1) = \int_{\mathbb{R}^{m-1}} \int_{-\infty}^{\infty} \frac{t}{\sqrt{t^2 + \|z\|_2^2}} \phi_1(t - \kappa) \phi_{m-1}(z) dt dz.$$

Setting $h(t, z) = t/\sqrt{t^2 + \|z\|_2^2}$, we have that

$$\begin{aligned} \int_{-\infty}^{\infty} h(t, z) \phi_1(t - \kappa) dt &= \int_{-\infty}^0 h(t, z) \phi_1(t - \kappa) dt + \int_0^{\infty} h(t, z) \phi_1(t - \kappa) dt \\ &= \int_{\infty}^0 h(t, z) \phi_1(-t - \kappa) dt + \int_0^{\infty} h(t, z) \phi_1(t - \kappa) dt \\ &= \int_{\infty}^0 h(t, z) \phi_1(t + \kappa) dt + \int_0^{\infty} h(t, z) \phi_1(t - \kappa) dt \\ &= \int_0^{\infty} h(t, z) [\phi_1(t - \kappa) - \phi_1(t + \kappa)] dt. \end{aligned} \tag{19}$$

Now, for any $t \geq 0$,

$$\begin{aligned} \phi_1(t - \kappa) - \phi_1(t + \kappa) &= \frac{e^{-(t^2 + \kappa^2)/2}}{\sqrt{2\pi}} (e^{t\kappa} - e^{-t\kappa}) \\ &= \frac{e^{-(t^2 + \kappa^2)/2}}{\sqrt{2\pi}} \tanh(t\kappa) (e^{t\kappa} + e^{-t\kappa}) \\ &\geq \frac{e^{-(t^2 + \kappa^2)/2}}{\sqrt{2\pi}} \tanh(t\kappa) e^{t\kappa} = \tanh(t\kappa) \phi_1(t - \kappa), \end{aligned}$$

hence

$$\begin{aligned}
\int_{\mathbb{R}^m} \left\langle \frac{y}{\|y\|_2}, -e_1 \right\rangle d\Phi(y + \kappa e_1) &\geq \int_{\mathbb{R}^{m-1}} \int_0^\infty h(t, z) \tanh(t\kappa) \phi_1(t - \kappa) \phi_{m-1}(z) dt dz \\
&\geq \int_{\|z\|_2 \leq R} \int_1^\infty h(t, z) \tanh(t\kappa) \phi_1(t - \kappa) \phi_{m-1}(z) dt dz \\
&\geq \frac{\tanh(\kappa)}{\sqrt{1+R^2}} \int_{\|z\|_2 \leq R} \phi_{m-1}(z) dz \int_1^\infty \phi_1(t - \kappa) dt \\
&\geq \frac{0.15 \tanh(\kappa)}{\sqrt{1+R^2}} \int_{\|z\|_2 \leq R} \phi_{m-1}(z) dz,
\end{aligned}$$

where we have use the inequality $h(t, z) \geq (1 + R^2)^{-1/2}$ whenever $\|z\|_2^2 \leq R$ and $t \geq 1$, and $1 - \Phi(1) > 0.15$. Finally, a well-known bound states that if Y has χ_{m-1}^2 distribution, then for all $t > 0$

$$\mathbb{P}\left(\frac{Y}{m-1} - 1 > t\right) \leq e^{-(m-1)t^2/8}.$$

For $R^2 := m - 1 + 2\sqrt{(m-1)\ln 4}$, it implies that

$$\int_{\|z\|_2 \leq R} \phi_{m-1}(z) dz = \mathbb{P}(Y \leq R^2) = \mathbb{P}\left(\frac{Y}{m-1} - 1 \leq 2\sqrt{\frac{\log 4}{m-1}}\right) \geq 1/2,$$

which concludes the proof.

4.8. Proof of Lemma 1.

It is a simple numerical fact that whenever $g(n) + \sqrt{\frac{s}{k}} \leq 0.33$, $\zeta(n, s) \leq 1$ (indeed, this follows since $\Phi(1) \simeq 0.8413 > 0.83$). Since $e^{-y^2/2} \geq 1 - \frac{y^2}{2}$, we have

$$\sqrt{2\pi} \left(g(n) + \sqrt{\frac{s}{k}} \right) = \int_0^{\zeta(n,s)} e^{-y^2/2} dy \geq \zeta(n, s) - \frac{1}{6} \zeta^3(n, s) \geq \frac{5}{6} \zeta(n, s), \quad (20)$$

where the last inequality follows since $\zeta(n, s) \leq 1$. Equation (20) implies that $\zeta(n, s) \leq \frac{6}{5} \sqrt{2\pi} \left(g(n) + \sqrt{\frac{s}{k}} \right)$. Proceeding again as in (20), we see that

$$\begin{aligned}
\sqrt{2\pi} \left(g(n) + \sqrt{\frac{s}{k}} \right) &\geq \zeta(n, s) - \frac{1}{6} \zeta^3(n, s) \\
&\geq \zeta(n, s) - \frac{12\pi}{25} \left(g(n) + \sqrt{s/k} \right)^2 \zeta(n, s) \geq \zeta(n, s) \left(1 - 1.51 \left(g(n) + \sqrt{s/k} \right)^2 \right),
\end{aligned}$$

hence $\zeta(n, s) \leq \frac{\sqrt{2\pi}}{1 - 1.51 \left(g(n) + \sqrt{s/k} \right)^2} \left(g(n) + \sqrt{s/k} \right)$, and result follows.

A. Supplementary results.

LEMMA 5. *Inequality $\tanh(x) \geq x \left(\frac{1+x}{1+x+x^2} \right)$ holds for all $x \geq 0$. Moreover, if $\tanh(x) \leq 1/2$ and $x \geq 0$, then $\tanh(x) \geq 0.83x$.*

PROOF. Since $e^x \geq 1 + x + \frac{x^2}{2}$ for all $x \geq 0$,

$$\tanh(x) = 1 - \frac{2}{1 + e^{2x}} \geq 1 - \frac{1}{1 + x + x^2} = x \left(\frac{1 + x}{1 + x + x^2} \right).$$

Note that $f(x) = \frac{1+x}{1+x+x^2}$ is decreasing on $[0, \infty)$. Whenever $\tanh(x) \leq 1/2$, $x \leq \frac{\log 3}{2} \leq 0.55$, hence $\tanh(x) \geq 0.83x$.

References

- Alon, N., Matias, Y. and Szegedy, M. (1996) The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, 20–29. ACM.
- Arcones, M. A. (1996) The Bahadur-Kiefer representation for U-quantiles. *The Annals of Statistics*, **24**, 1400–1422.
- Battey, H., Fan, J., Liu, H., Lu, J. and Zhu, Z. (2015) Distributed estimation and inference with statistical guarantees. *arXiv preprint arXiv:1509.05457*.
- Bentkus, V. (2003) On the dependence of the berry–esseen bound on dimension. *Journal of Statistical Planning and Inference*, **113**, 385–402.
- Berry, A. C. (1941) The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, **49**, 122–136.
- Bickel, P. J. et al. (1965) On some robust estimates of location. *The Annals of Mathematical Statistics*, **36**, 847–858.
- Bubeck, S., Cesa-Bianchi, N. and Lugosi, G. (2013) Bandits with heavy tail. *IEEE Transactions on Information Theory*, **59**, 7711–7717.
- Catoni, O. (2012) Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, vol. 48, 1148–1185. Institut Henri Poincaré.
- Cheng, G. and Shang, Z. (2015) Computational limits of divide-and-conquer method. *arXiv preprint arXiv:1512.09226*.
- Devroye, L., Lerasle, M., Lugosi, G., Oliveira, R. I. et al. (2016) Sub-Gaussian mean estimators. *The Annals of Statistics*, **44**, 2695–2725.
- Duchi, J. C., Jordan, M. I., Wainwright, M. J. and Zhang, Y. (2014) Optimality guarantees for distributed statistical estimation. *arXiv preprint arXiv:1405.0782*.

- Dudley, R. M. (1978) Central limit theorems for empirical measures. *The Annals of Probability*, 899–929.
- Esseen, C.-G. (1942) *On the Liapounoff limit of error in the theory of probability*. Almqvist & Wiksell.
- Fan, J., Han, F. and Liu, H. (2014) Challenges of Big Data analysis. *National science review*, **1**, 293–314.
- Fan, J., Wang, D., Wang, K. and Zhu, Z. (2017) Distributed estimation of principal eigenspaces. *arXiv preprint arXiv:1702.06488*.
- Giné, E. and Nickl, R. (2015) *Mathematical foundations of infinite-dimensional statistical models*, vol. 40. Cambridge University Press.
- Haussler, D. (1995) Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, **69**, 217–232.
- Hodges, J. L. and Lehmann, E. L. (1963) Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, 598–611.
- Hoeffding, W. (1948) A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 293–325.
- (1963) Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, **58**, 13–30.
- Hsu, D. and Sabato, S. (2013) Loss minimization and parameter estimation with heavy tails. *arXiv preprint arXiv:1307.1827*.
- (2016) Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, **17**, 1–40.
- IBM (2015) What is Big Data? <https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>.
- Jerrum, M. R., Valiant, L. G. and Vazirani, V. V. (1986) Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, **43**, 169–188.
- Joly, E., Lugosi, G. and Oliveira, R. I. (2016) On the estimation of the mean of a random vector. *arXiv preprint arXiv:1607.05421*.
- Jordan, M. (2013) On statistics, computation and scalability. *Bernoulli*, **19**, 1378–1390.
- Lee, J. D., Sun, Y., Liu, Q. and Taylor, J. E. (2015) Communication-efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*.
- Lehmann, E. L. and D’Abrera, H. J. (2006) *Nonparametrics: statistical methods based on ranks*. Springer New York.

- Lerasle, M. and Oliveira, R. I. (2011) Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*.
- Li, C., Srivastava, S. and Dunson, D. B. (2016) Simple, scalable and accurate posterior interval estimation. *arXiv preprint arXiv:1605.04029*.
- Liang, Y., Balcan, M.-F. F., Kanchanapally, V. and Woodruff, D. (2014) Improved distributed Principal Component Analysis. In *Advances in Neural Information Processing Systems*, 3113–3121.
- Lugosi, G. and Mendelson, S. (2017) Sub-Gaussian estimators of the mean of a random vector. *arXiv preprint arXiv:1702.00482*.
- Mcdonald, R., Mohri, M., Silberman, N., Walker, D. and Mann, G. S. (2009) Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems*, 1231–1239.
- Minsker, S., Srivastava, S., Lin, L. and Dunson, D. B. (2014) Robust and scalable Bayes via a median of subset posterior measures. *arXiv preprint arXiv:1403.2660*.
- Minsker, S. a. (2015) Geometric median and robust estimation in Banach spaces. *Bernoulli*, **21**, 2308–2335.
- Nemirovski, A. and Yudin, D. (1983) *Problem complexity and method efficiency in optimization*. John Wiley & Sons Inc.
- Pinelis, I. (2016) Optimal-order bounds on the rate of convergence to normality for maximum likelihood estimators. *arXiv preprint arXiv:1601.02177*.
- Pollard, D. (2000) Asymptopia: an exposition of statistical asymptotic theory. Available at <http://www.stat.yale.edu/~pollard/Books/Asymptopia>.
- Rosenblatt, J. D. and Nadler, B. (2016) On the optimality of averaging in distributed statistical learning. *Information and Inference*, **5**, 379–404.
- Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. and McCulloch, R. E. (2016) Bayes and big data: the consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, **11**, 78–88.
- Shang, Z. and Cheng, G. (2015) A Bayesian splitotic theory for nonparametric models. *arXiv preprint arXiv:1508.04175*.
- Shevtsova, I. (2011) On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands. *arXiv preprint arXiv:1111.6554*.
- Small, C. (1990) A survey of multidimensional medians. *International Statistical Review*, **58**, 263–277.
- Talagrand, M. (2005) *The generic chaining*. Springer.

- Tukey, J. and Harris, T. (1946) Sampling from contaminated distributions. *Ann. Math. Statist.* **17**501.
- van der Vaart, A. W. (1998) *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge University Press.
- van der Vaart, A. W. and Wellner, J. A. (1996) *Weak convergence and empirical processes*. Springer Series in Statistics. New York: Springer-Verlag.
- Zhang, Y., Duchi, J. and Wainwright, M. (2013) Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, 592–617.
- Zhang, Y., Wainwright, M. J. and Duchi, J. C. (2012) Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, 1502–1510.
- Zinkevich, M., Weimer, M., Li, L. and Smola, A. J. (2010) Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, 2595–2603.

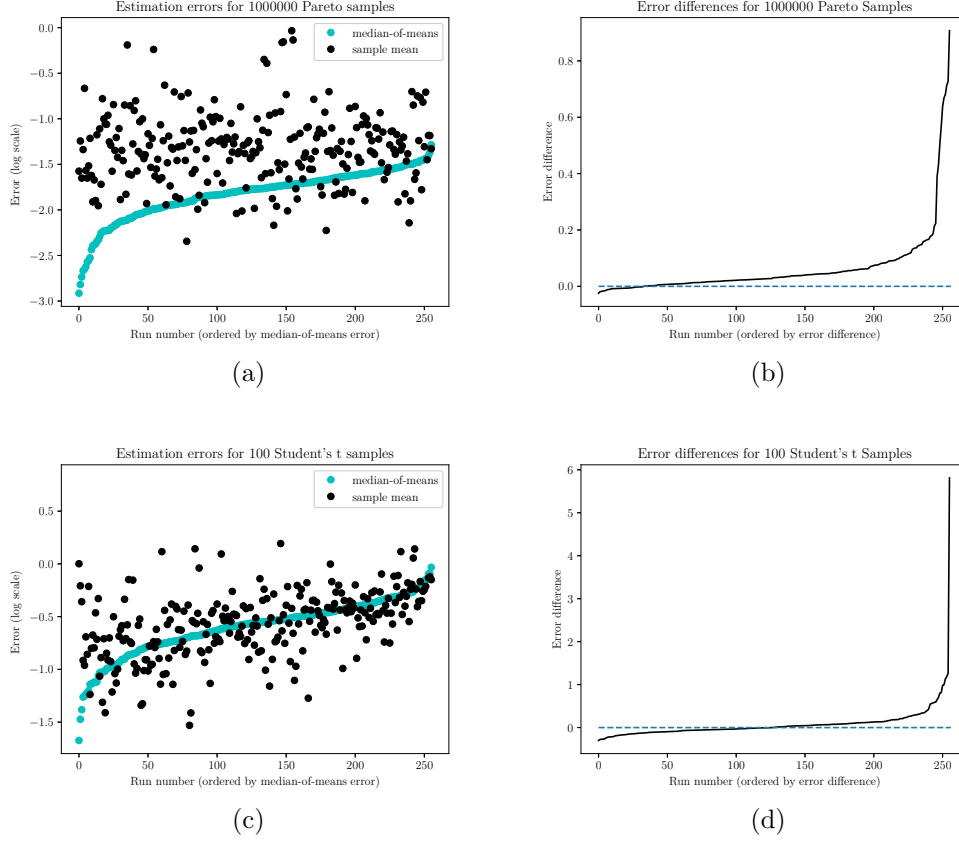


Fig. 2: Comparison of errors corresponding to the median-of-means and sample mean estimator over 256 runs of the experiment. In (a) the sample of size $N = 10^6$ consists of i.i.d. random vectors in \mathbb{R}^2 with independent Pareto-distributed entries possessing only 2.1 moments. Each run computes the (geometric) median-of-means estimator using partition into $k = 1000$ groups, as well as the usual sample mean. In (b), the ordered differences between the error of the sample mean and the median-of-means over all 256 runs illustrates robustness. Positive error differences in (b) indicate lower error for the median-of-means, and negative error differences occur when the sample mean provided a better estimate.

Images (c) and (d) illustrate a similar experiment that was performed for two-dimensional random vectors with independent entries with Student's t-distribution with 2 degrees of freedom. In this case, the sample size is $N = 100$ and the number of groups is $k = 10$.